Code No: RT32052 **R13** SET - 1

III B. Tech II Semester Regular/Supplementary Examinations, April -2018

DATA WARE HOUSING AND MINING

(Common to Computer Science Engineering and Information Technology)

Time: 3 hours Max. Marks: 70

Note: 1. Question Paper consists of two parts (**Part-A** and **Part-B)**
2. Answering the question in **Part-A** is compulsory
3. Answer any **THREE** Questions from **Part-B**

*****

PART –A

1	a)	Why data mining is required?	[3M]
	b)	With an example, justify the need of data Integration?	[4M]
	c)	Compare and contrast ROLAP versus MOLAP.	[3M]
	d)	Justify the need of attribute splitting rules? Where one is used?	[4M]
	e)	What is pruning? Why support-based pruning is required?	[4M]
	f)	Why clustering called unsupervised classification?	[4M]

PART –B

2	a)	What is the difference between discrimination and classification? Between characterization and clustering? Between classification and prediction? For each of these pairs of tasks, how are they similar?	[8M]
	b)	Briefly describe data mining functionalities.	[8M]
3	a)	What is Preprocessing? Why we need to preprocess the data? Briefly describe the forms of data preprocessing.	[8M]
	b)	What is data reduction? Describe the strategies for data reduction.	[8M]
4	a)	Briefly describe the available processes for data cube materialization.	[8M]
	b)	With an example, describe the Efficient Data Cube Computation.	[8M]
5	a)	What is attribute selection measure? Briefly describe the attribute selection measures for decision tree induction.	[8M]
	b)	With an example, describe the classification by decision tree induction.	[8M]
6	a)	Consider the following set of frequent 3-itemsets: {1, 2, 3}, {1, 2, 4}, {1, 2, 5}, {1, 3, 4}, {1, 3, 5}, {2, 3, 4}, {2, 3, 5}, {3, 4, 5}. Assume that there are only five items in the data set. i) List all candidate 4-itemsets obtained by the candidate generation procedure in *Apriori*. ii) List all candidate 4-itemsets obtained by a candidate generation procedure using the $F_{k-1} \times F_1$ merging strategy.	[8M]
	b)	Briefly describe Apriori algorithm for frequent itemset generation.	[8M]
7	a)	How to handle empty clusters and outliers in k-means clustering?	[8M]
	b)	Compare and contrast K-means clustering Vs Hierarchical clustering.	[8M]

*****

Code No: RT32052 **R13** SET - 2

III B. Tech II Semester Regular/Supplementary Examinations, April -2018

DATA WARE HOUSING AND MINING

(Common to Computer Science Engineering and Information Technology)

Time: 3 hours Max. Marks: 70

Note: 1. Question Paper consists of two parts (**Part-A** and **Part-B)**
2. Answering the question in **Part-A** is compulsory
3. Answer any **THREE** Questions from **Part-B**

*****

PART –A

1	a)	What are challenges of data mining?	[3M]
	b)	Justify the need of data reduction?	[4M]
	c)	Briefly describe key features of data warehouse.	[3M]
	d)	How entropy is used in classification?	[4M]
	e)	Why confidence-based pruning is required?	[4M]
	f)	Would the cosine measure be the appropriate similarity measure to use with K-means clustering for time series data? Why or why not?	[4M]

PART -B

2	a)	What are the major challenges of mining a huge amount of data (such as billions of tuples) in comparison with mining a small amount of data (such as a few hundred tuple data set)?	[8M]
	b)	Describe the differences between Operational Database Systems and Data Warehouses.	[8M]
3	a)	What is descriptive data summarization? Why descriptive data summarization is used? What is dispersion? Describe measures for Measuring the Dispersion of Data.	[8M]
	b)	What is attribute subset selection? Describe heuristic methods of attribute subset selection.	[8M]
4	a)	Describe various schemes used for the design of multidimensional data model.	[8M]
	b)	With an example, describe indexing OLAP data using bitmap indices.	[8M]
5	a)	Briefly describe the measures for selecting the bet split.	[6M]
	b)	What is cross validation? With an example, describe how cross validation can be used for evaluating the performance of a classification model.	[10M]

Code No: RT32052 **R13** SET - 2

6 a) [8M]

Transaction ID	Items Bought
1	{Milk, Beer, Diapers}
2	{Bread, Butter, Milk}
3	{Milk, Diapers, Cookies}
4	{Bread, Butter, Cookies}
5	{Beer, Cookies, Diapers}
6	{Milk, Diapers, Bread, Butter}
7	{Bread, Butter, Diapers}
8	{Beer, Diapers}
9	{Milk, Diapers, Bread, Butter}
10	{Beer, Cookies}

Consider the market basket transactions shown in the above table:
i) What is the maximum number of association rules that can be extracted from this data (including rules that have zero support)?
ii) What is the maximum size of frequent itemsets that can be extracted (assuming minsup > 0)?

b) Briefly describe the factors that can affect the computational complexity of Apriori algorithm. [8M]

7 a) For your own data, describe step-by-step process of bisecting k-means clustering. In what way bisecting k-means clustering is different from basic k-means clustering. [8M]

b) Compare and contrast DBSCAN clustering Vs Hierarchical clustering. [8M]

*****

Code No: RT32052 **R13** SET - 3

III B. Tech II Semester Regular/Supplementary Examinations, April -2018

DATA WARE HOUSING AND MINING

(Common to Computer Science Engineering and Information Technology)

Time: 3 hours Max. Marks: 70

Note: 1. Question Paper consists of two parts (**Part-A** and **Part-B)**
2. Answering the question in **Part-A** is compulsory
3. Answer any **THREE** Questions from **Part-B**

*****

PART –A

1	a)	What are the task primitives of data mining?	[3M]
	b)	Justify the need of Data Discretization?	[4M]
	c)	What is partial materialization? Why it is required?	[4M]
	d)	How information gain is used in classification?	[3M]
	e)	What are item sets? How can one reduce the number of candidate item sets?	[4M]
	f)	Total SSE is the sum of the SSE for each separate attribute. What does it mean if the SSE for one variable is low for all clusters?	[4M]

PART -B

2	a)	What is data characterization and data discrimination? Why these are required?	[8M]
	b)	What is transactional database? Describe any five advanced database systems.	[8M]
3	a)	What is Data cleaning? Describe the techniques for handling missing values and noisy data.	[8M]
	b)	What is concept hierarchy generation? Describe Concept Hierarchy Generation for Categorical Data.	[8M]
4	a)	With an example, describe the usage of composite join indices.	[8M]
	b)	What is query driven approach and what is data driven approach? How these can be utilized while building data warehouses?	[8M]
5	a)	Briefly describe impurity measures that are used for selecting the best split and compare them for binary classification problems.	[8M]
	b)	Describe with an example, how model over-fitting can happen due to the presence of noise?	[8M]

Code No: RT32052 **R13** SET - 3

6 a) [8M]

Transaction ID	Items Bought
1	{Milk, Beer, Diapers}
2	{Bread, Butter, Milk}
3	{Milk, Diapers, Cookies}
4	{Bread, Butter, Cookies}
5	{Beer, Cookies, Diapers}
6	{Milk, Diapers, Bread, Butter}
7	{Bread, Butter, Diapers}
8	{Beer, Diapers}
9	{Milk, Diapers, Bread, Butter}
10	{Beer, Cookies}

Consider the market basket transactions shown in the above table:
(i) Write an expression for the maximum number of size-3 itemsets that can be derived from this data set.
(ii) Find an item set (of size 2 or larger) that has the largest support.

b) Briefly describe the ways to reduce the computational complexity of frequent item set generation. [4M]

c) What is candidate generation? List the requirements for an effective candidate generation. [4M]

7 a) For a suitable data, describe the step-by-step process of k-means clustering. [8M]

b) What is DBSCAN? For which situation you suggest the usage of DBSCAN clustering? [8M]

*****

Code No: RT32052 **R13** SET - 4

III B. Tech II Semester Regular/Supplementary Examinations, April -2018

DATA WARE HOUSING AND MINING

(Common to Computer Science Engineering and Information Technology)

Time: 3 hours Max. Marks: 70

Note: 1. Question Paper consists of two parts (**Part-A** and **Part-B)**
2. Answering the question in **Part-A** is compulsory
3. Answer any **THREE** Questions from **Part-B**

*****

PART –A

1	a)	What is the need of data warehouse?	[3M]
	b)	With an example, justify the need of data Transformation?	[4M]
	c)	Justify the need of bit map indexing and join indexing.	[4M]
	d)	What are attribute selection measures? Why they require?	[4M]
	e)	What are maximal frequent item sets? Why they require?	[4M]
	f)	Justify the need of graph-based clustering?	[3M]

PART -B

2	a)	Describe three challenges to data mining regarding data mining methodology and user interaction issues.	[8M]
	b)	Present an example where data mining is crucial to the success of a business. What data mining functionalities does this business need (e.g., think of the kinds of patterns that could be mined)? Can such patterns be generated alternatively by data query processing or simple statistical analysis?	[8M]
3	a)	Why correlation analysis is useful? How correlation coefficient is computed?	[8M]
	b)	Suppose a group of 12 sales price records has been sorted as follows: 5, 10, 11, 13, 15, 35, 50, 55, 72, 92, 204, 215; Partition them into three bins by each of the following methods: (i) equal-frequency (equidepth) partitioning; (ii) equal-width partitioning	[8M]
4	a)	What is OLAM? Why OLAM is important? Describe the OLAM architecture.	[8M]
	b)	Compare and contrast OLTP Vs OLAP.	[8M]
5	a)	What is gain ratio? Briefly describe splitting of continuous attributes.	[8M]
	b)	Describe with an example, how model over-fitting can happen due to lack of representation samples?	[8M]

Code No: RT32052 **R13** SET - 4

6 a) [8M]

Transaction ID	Items Bought
1	{Milk, Beer, Diapers}
2	{Bread, Butter, Milk}
3	{Milk, Diapers, Cookies}
4	{Bread, Butter, Cookies}
5	{Beer, Cookies, Diapers}
6	{Milk, Diapers, Bread, Butter}
7	{Bread, Butter, Diapers}
8	{Beer, Diapers}
9	{Milk, Diapers, Bread, Butter}
10	{Beer, Cookies}

Consider the market basket transactions shown in the above table:
(i) What is the maximum size of frequent itemsets that can be extracted (assuming minsup > 0)?
(ii) Find a pair of items, *a* and *b*, such that the rules {a} → {b} and {b} → {a} have the same confidence.

b) Briefly describe the relation among frequent, maximal frequent and closed frequent item sets. [8M]

7 a) Highlight strengths and weaknesses of k-means clustering algorithm. [8M]

b) With an example, briefly describe the construction of dendograms. [8M]

*****